

# Estimating the coefficients of a mixture of two linear regressions by expectation maximization

Jason M. Klusowski      Dana Yang      W. D. Brinda

Department of Statistics  
Yale University  
New Haven, CT, USA

Email: {jason.klusowski, xiaoqian.yang, william.brinda}@yale.edu

## Abstract

We give convergence guarantees for estimating the coefficients of a symmetric mixture of two linear regressions by expectation maximization (EM). In particular, if the initializer has a large cosine angle with the population coefficient vector and the signal to noise ratio (SNR) is large, a sample-splitting version of the EM algorithm converges to the true coefficient vector with high probability. Here “large” means that each quantity is required to be at least a universal constant. Finally, we show that the population EM operator is not globally contractive by characterizing a region where it fails. Interestingly, our analysis borrows from tools used in the problem of estimating the centers of a symmetric mixture of two Gaussians by EM.

## 1 Introduction

The Expectation-Maximization (EM) algorithm is a widely used technique for parameter estimation. It is an iterative procedure that monotonically increases the likelihood. When the likelihood is not concave, it is well known that EM can converge to a non-global optimum. However, recent work has side-stepped the question of whether EM reaches the likelihood maximizer. Instead directly working out statistical guarantees on its loss. These explorations have identified regions of initialization for which the EM estimate approaches the true parameter in probability, assuming the model is well-specified.

This line of research was spurred by [1] which established general conditions for which a ball centered at the true parameter would be a basin of attraction for the population version of the EM operator. For a large enough sample size, the difference (in that ball) between the sample EM operator and the population EM operator can be bounded such that the EM estimate approaches the true parameter with high probability. That bound is the sum of two terms with distinct interpretations. There is an *algorithmic convergence* term  $\kappa^t \|\theta^0 - \theta^*\|$  for initializer  $\theta^0$ , truth  $\theta^*$ , and some modulus of contraction  $\kappa \in (0, 1)$ ; this comes from the analysis of the population EM operator. The second term captures *statistical convergence* and is proportional to the supremum norm of  $M - M_n$ , the difference between the population and sample EM operators, over the ball. This result is also shown for a “sample-splitting” version of EM, where the sample is partitioned into batches and each batch governs a single step of the algorithm.

That article also detailed three specific simple models in which their analysis is easily seen to apply: symmetric mixture of two spherical Gaussians, symmetric mixture of two linear models with

Gaussian covariates and error, and linear regression with data missing completely at random.

The performance of EM for their first example, a symmetric mixture of two spherical Gaussians, has since received further attention. [9] showed that the intersection of a suitable half-space and ball about the origin is also a basin of attraction for the population EM in that model when the component means are separated well enough relative to the noise. Exact probabilistic bounds on the error of the EM estimate were also derived when the initializer is in the region. The authors also proposed a random initialization strategy that has a high probability of finding the basin of attraction when the component means are well-separated as  $\sqrt{d \log d}$ . Concurrently, [14] revealed that the entirety of  $\mathbb{R}^d$  (except the hyperplane perpendicular to  $\theta^*$ ) is a basin of attraction for the population EM operator (in addition to asymptotic consistency of the empirical iterates). Subsequently in [5], a more explicit expression for the contraction constant and its dependence on the initializer was obtained through an elegant argument. These refinements can be used to improve the separation requirement in [9] to  $\sqrt{d}$ .

The second example of [1], the symmetric mixture of two linear models with Gaussian covariates and error, can be seen as a generalization of the symmetric mixture. This model, also known as Hierarchical Mixture of Experts (HME) in the machine learning community [8], has drawn recent attention (e.g. [4], [15], [3], [16], [11]). The analysis of the two-mixture case was generalized to arbitrary multiple components in [16], but initialization is still required to be in a ball around each of the true coefficient vectors.

Our purpose here is to follow up the analysis of [1] by proving a larger basin of attraction for the mixture of two linear models and by establishing an exact probabilistic bound on the error of the sample-splitting EM estimate when the initializer falls in the specified region.

In Section 2, we explain the model and derive a basin of attraction for the population version of the EM operators and also show that it is not contractive in certain regions of  $\mathbb{R}^d$ . Section 3 looks at the behavior of the sample-splitting EM operator in this region and proves statistical guarantees. Section 4 considers a more general model that doesn't require symmetry. We point out that estimation for that model can be handled by an estimator related to the symmetric case's EM; this estimator essentially inherits the statistical guarantees derived for EM in the symmetric case. Finally, the more technical proofs are in Appendix A.

## 2 The population EM operator

Let data  $(X_i, Y_i)_{i=1}^n$  be *i.i.d.* with  $X_i \sim N(0, I_d)$  and

$$Y_i = R_i \langle \theta^*, X_i \rangle + \varepsilon_i$$

where  $\varepsilon_i \sim N(0, \sigma^2)$ ,  $R_i \sim \text{Rademacher}$ , and  $X_i, \varepsilon_i, R_i$  are independent of each other. In other words, each predictor variable is normal, and the response is centered at either the  $\theta^*$  or  $-\theta^*$  linear combination of the predictor. The two classes are equally probable, and the label of each observation is unknown. We seek to estimate  $\theta^*$  (or  $-\theta^*$ , which produces the same model distribution).

The likelihood function is multi-model, and direct maximization is infeasible. The EM algorithm has been used to estimate the model coefficients [8], and simulation studies have shown that it has desirable empirical performance [6], [13], [7]. The EM operator for estimating  $\theta^*$  (see [1, page 6] for a derivation) is

$$M_n(\theta) = \left( \frac{1}{n} \sum X_i X_i^T \right)^{-1} \left[ \frac{1}{n} \sum (2\phi(Y_i \langle \theta, X_i \rangle / \sigma^2) - 1) X_i Y_i \right] \quad (1)$$

where  $\phi(t) = \frac{1}{1+e^{-2t}}$  is a horizontally stretched logistic sigmoid. The population EM operator replaces sample averages with expectations, thus

$$M(\theta) = 2\mathbb{E}[\phi(Y\langle\theta, X\rangle/\sigma^2)XY]. \quad (2)$$

Conveniently, this estimation can be reduced to the  $\sigma = 1$  case. If we divide each response datum by  $\sigma$ :

$$Y_i/\sigma = R_i\langle\theta^*/\sigma, X_i\rangle + \varepsilon_i/\sigma,$$

the unknown parameter to estimate becomes  $\theta^*/\sigma$ , and the noise has variance 1. Inspection of (1) and (2) confirms that the EM operators for the new problem are equal to  $1/\sigma$  times the EM operators for the original problem. For instance, denoting the population EM operator of the new problem by  $\widetilde{M}$ ,

$$\begin{aligned} \widetilde{M}(\theta/\sigma) &= 2\mathbb{E}[\phi((Y/\sigma)\langle(\theta/\sigma), X\rangle)X(Y/\sigma)] \\ &= \frac{2}{\sigma}\mathbb{E}[\phi(Y\langle\theta, X\rangle/\sigma^2)XY] \\ &= \frac{1}{\sigma}M(\theta) \end{aligned}$$

The transformed problem's error is easily related to the original problem's error:

$$\begin{aligned} \|\widetilde{M}(\theta/\sigma) - \theta^*/\sigma\| &= \|\frac{1}{\sigma}M(\theta) - \theta^*/\sigma\| \\ &= \frac{1}{\sigma}\|M(\theta) - \theta^*\| \end{aligned}$$

Thus, in the general case, the estimation error is exactly  $\sigma$  times the estimation error of the normalized problem. We use this observation to simplify the proof of Lemma 1, while stating our results for general  $\sigma$ .

In [1], it was shown that if the EM algorithm is initialized in a ball around  $\theta^*$  with radius proportional  $\theta^*$ , the EM algorithm converges with high probability. The purpose of this paper is to relax these conditions and show that if the cosine angle between  $\theta^*$  and the initializer is not too small, the EM algorithm also converges. We also simplify the analysis, using only elementary facts about multivariate normal distributions. This improvement is manifested in the set containment

$$\{\theta : \|\theta - \theta^*\| \leq \sqrt{1 - \rho^2}\|\theta^*\|\} \subseteq \{\theta : \langle\theta, \theta^*\rangle \geq \rho\|\theta\|\|\theta^*\|\}, \quad \rho \in [0, 1],$$

since for all  $\theta$  in the set on the left side,

$$\begin{aligned} \langle\theta, \theta^*\rangle &= \frac{1}{2}(\|\theta\|^2 + \|\theta^*\|^2 - \|\theta - \theta^*\|^2) \\ &\geq \frac{1}{2}(\|\theta\|^2 + \rho^2\|\theta^*\|^2) \\ &\geq \rho\|\theta\|\|\theta^*\|. \end{aligned}$$

The authors of [1] required the initializer  $\theta^0$  to be at most  $\|\theta^*\|/32$  away from  $\theta^*$ , which translates to  $\rho \geq \sqrt{1 - (1/32)^2}$ , while we only need  $\rho \geq \sqrt{1 - 0.35^2}$ . We would also like to point out that the set of initializers we allow is unbounded.

We will also show how the analysis relates to the one-dimensional mixture of two Gaussians by exploiting the self-consistency property of its population EM operator.

Let  $\theta_0$  be the unit vector in the direction of  $\theta$  and let  $\theta_0^\perp$  be the unit vector that belongs to the hyperplane spanned by  $\{\theta^*, \theta\}$  and orthogonal to  $\theta$  (i.e.  $\theta_0^\perp \in \text{span}\{\theta, \theta^*\}$  and  $\langle\theta, \theta_0^\perp\rangle = 0$ ). Let  $\theta^\perp = \|\theta\|\theta_0^\perp$ . We will later show that  $M(\theta)$  belongs to  $\text{span}\{\theta, \theta^*\}$ , as in Fig. 1. Denote the angle between  $\theta^*$  and  $\theta_0$  as  $\alpha$ , with  $\|\theta^*\|\cos\alpha = \langle\theta_0, \theta^*\rangle$  and  $\rho = \cos\alpha$ . As we will see from the following results, as long as  $\cos\alpha$  is not too small,  $M(\theta)$  is a contracting operation that is always closer to the truth  $\theta^*$  than  $\theta$ .

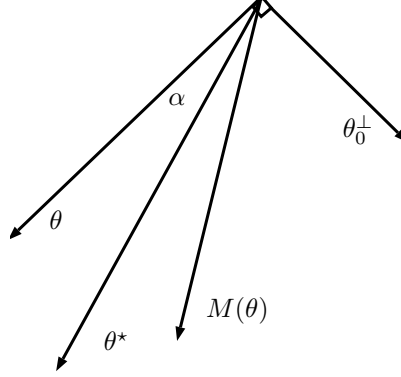


Figure 1: The population EM operator  $M(\theta)$  lies in the space spanned by  $\theta$  and  $\theta^*$ . The unit vector  $\theta_0^\perp$  lies in the space spanned by  $\theta$  and  $\theta^*$  and is perpendicular to  $\theta$ . The vector  $\theta$  forms an angle  $\alpha$  with  $\theta^*$ .

**Lemma 1.** For any  $\theta$  in  $\mathbb{R}^d$  with  $\langle \theta, \theta^* \rangle \geq 0$ ,

$$\|M(\theta) - \theta^*\| \leq \sqrt{\kappa} \sqrt{1 + 4 \left( \frac{|\langle \theta^\perp, \theta^* \rangle| + \sigma^2}{\langle \theta, \theta^* \rangle} \right)^2} \|\theta - \theta^*\|, \quad (3)$$

where

$$\kappa = \max \left\{ \left( 1 - \frac{|\langle \theta_0, \theta^* \rangle|^2}{\sigma^2 + \|\theta^*\|^2} \right)^{1/2}, \left( \frac{\sigma^2}{\sigma^2 + \langle \theta, \theta^* \rangle} \right)^{1/2} \right\} \leq 1. \quad (4)$$

As we will see, this constant  $\kappa$  is closely related to the contraction constant  $\gamma$  of the operator  $M(\theta)$ .

If we write the signal to noise ratio as  $\eta = \|\theta^*\|/\sigma$  and use the fact that  $\|\theta^*\| \cos \alpha = \langle \theta_0, \theta^* \rangle$ , the contractivity constant can be written as

$$\max \left\{ \left( 1 - \frac{\eta^2 \cos^2 \alpha}{1 + \eta^2} \right)^{1/4}, \left( \frac{1}{1 + (\|\theta\|/\sigma) \eta \cos \alpha} \right)^{1/4} \right\} \sqrt{1 + 4 \left( \tan \alpha + \frac{\sigma}{\|\theta\|} \frac{1}{\eta \cos \alpha} \right)^2}. \quad (5)$$

**Remark 1.** If  $\|\theta\| \geq 3\sigma$ ,  $\|\theta^*\| \geq 4\sigma$  and  $\cos \alpha \geq \sqrt{1 - 0.35^2}$ , the quantity (5) is bounded by a universal constant  $\gamma < 1/4$ , implying the population EM operator  $\theta^{t+1} \leftarrow M(\theta^t)$  converges to the truth  $\theta^*$  exponentially fast.

This next theorem shows that  $M(\theta)$  may not be contractive for all  $\theta$ . In contrast, it is known that the population EM operator for a symmetric mixture of two Gaussians is globally contractive [5], [14].

**Theorem 1.** There are points  $\theta$  satisfying  $\langle \theta, \theta^* \rangle = 0$  such that

$$\|M(\theta) - \theta^*\| > \|\theta - \theta^*\|.$$

### 3 The sample EM operator

As in [1], we analyze a sample-splitting version of the EM algorithm, where for an allocation of  $n$  samples and  $T$  iterations, we divide the data into  $T$  subsets of size  $\lfloor n/T \rfloor$ . We then perform the updates  $\theta^{t+1} \leftarrow M_{n/T}(\theta^t)$ , using a new subset of samples to compute  $M_{n/T}(\theta)$  at each iteration.

**Theorem 2.** Let  $\langle \theta^0, \theta^* \rangle > \rho \|\theta^0\| \|\theta^*\|$ ,  $3\sigma \leq \|\theta^0\| \leq R\sigma$ , and  $\|\theta^*\| \geq 4\sigma$  for  $\rho \in (\sqrt{1 - 0.35^2}, 1)$ . Suppose furthermore that  $n \geq cd \log(1/\delta)$  for  $\delta \in (0, 1)$  and some constant  $c = c(\rho, \sigma, \|\theta^*\|, R) \geq 1$ . Then there exists  $\gamma = \gamma(\rho, \sigma, \|\theta^*\|) \in (0, 1/4)$  such that the sample-splitting empirical EM iterates  $\{\theta^t\}_{t=1}^T$  based on  $n/T$  samples per step satisfy

$$\|\theta^t - \theta^*\| \leq \gamma^t \|\theta^0 - \theta^*\| + \frac{C \sqrt{\|\theta^*\|^2 + \sigma^2}}{1 - \gamma} \sqrt{\frac{dT \log(T/\delta)}{n}},$$

with probability at least  $1 - \delta$ .

We will prove this result at the end of the paper. The main aspect of the analysis lies in showing that  $M_n$  satisfies an invariance property:  $M_n(\mathcal{A}) \subseteq \mathcal{A}$ , where  $\mathcal{A}$  is the basin of attraction. The algorithmic error  $\gamma^t \|\theta^0 - \theta^*\|$  follows from Lemma 1 and the stochastic error  $\frac{C \sqrt{\|\theta^*\|^2 + \sigma^2}}{1 - \gamma} \sqrt{\frac{dT \log(T/\delta)}{n}}$  from the proof of Corollary 4 in [1].

**Remark 2.** Theorem 2 requires the initializer to have a good inner product with  $\theta^*$ . But how to initialize in practice? Initialization based on spectral [15], [3], [16] or Bayesian [13] methods can work well.

## 4 Without assuming symmetry

Without requiring symmetry, we can still derive statistical guarantees for a variant on the EM estimation procedure described above. In this section, we assume that data  $(X_i, Y_i)_{i=1}^n$  is i.i.d. with  $X_i \sim N(0, I_d)$  and

$$Y_i = \mathbb{1}\{R_i = 1\} \langle \theta_1^*, X_i \rangle + \mathbb{1}\{R_i = -1\} \langle \theta_2^*, X_i \rangle + \varepsilon_i$$

where  $\varepsilon_i \sim N(0, \sigma^2)$ ,  $R_i \sim \text{Rademacher}$ , and  $X_i, \varepsilon_i, R_i$  are independent of each other.

This time each model distribution is specified (uniquely up to class labels) by two parameters:  $\theta_1^*$  and  $\theta_2^*$ . Our previous analysis was for the restriction of this model to the slice in which  $\theta_2^* = -\theta_1^*$ .

Our first step is to reformulate the model as a shifted version of the symmetric case:

$$Y_i = R_i \langle \theta^*, X_i \rangle + \langle s, X_i \rangle + \varepsilon_i,$$

where  $\theta^* := (\theta_1^* - \theta_2^*)/2$  and the shift is  $s := (\theta_1^* + \theta_2^*)/2$ . The shift can be estimated by  $\hat{s} = \frac{1}{n} \sum_{i=1}^n X_i Y_i$  (or alternatively by  $(\frac{1}{n} \sum_{i=1}^n X_i X_i^T)^{-1} \hat{s}$ ) which concentrates around  $s$ . We construct a shifted version of the response vector and define an estimate for it:

$$\tilde{Y}_i := Y_i - \langle s, X_i \rangle \quad \text{and} \quad Y_i^{(s)} := Y_i - \langle \hat{s}, X_i \rangle$$

We use the symmetric model version of the EM algorithm on the approximately symmetric data  $(X_i, Y_i^{(s)})$  to define the estimator  $\hat{\theta}$  for  $\theta^*$ . The error incurred by the use of the estimated  $\hat{s}$  can be handled separately from the performance of EM on the truly symmetric  $(X_i, \tilde{Y}_i)$ , via the triangle inequality:

$$\|M_n(\theta, \underline{X}, \underline{Y}^{(s)}) - \theta^*\| \leq \|M_n(\theta, \underline{X}, \underline{Y}^{(s)}) - M_n(\theta, \underline{X}, \underline{\tilde{Y}})\| + \|M_n(\theta, \underline{X}, \underline{\tilde{Y}}) - \theta^*\|. \quad (6)$$

where each underlined letter represents the corresponding vector of  $n$  variables. Theorem 2 provides guarantees for good control on the second term of (6). The first term is small since the update procedure  $M_n$  is a smooth function of the data; it is of asymptotically smaller order than the second term. Finally, if desired, one can estimate the original parameters by  $\theta_1^t := \theta^t + \hat{s}$  and  $\theta_2^t := \hat{s} - \theta^t$ . The proof for the asymmetric case is below.

**Lemma 2.** *There exists constant  $D_1 > 0$ , such that*

$$\mathbb{P} \left\{ \|\hat{s} - s\| \leq D_1 \sqrt{\frac{d}{n} ((\|\theta_1^*\|^2 + \|\theta_2^*\|^2)/2 + \sigma^2) \log(1/\delta)} \right\} \geq 1 - \delta$$

for all  $\delta \in (0, 1)$ .

*Proof.* Denote  $\hat{\Sigma} = \frac{1}{n} X_i X_i^T$ . Recall that  $\hat{s} = \frac{1}{n} \sum X_i Y_i$ . We have

$$\begin{aligned} \|\hat{s} - s\| &= \left\| \frac{1}{n} \sum_{i=1}^n X_i (\langle s, X_i \rangle + R_i \langle \theta^*, X_i \rangle + \varepsilon_i) - s \right\| \\ &\leq \left\| (\hat{\Sigma} - I) s \right\| + \left\| \frac{1}{n} \sum_{i=1}^n R_i X_i X_i^T \theta^* \right\| + \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i \right\| \\ &\leq c \sqrt{\frac{d}{n} \log(1/\delta) (\|s\|^2 + \|\theta^*\|^2 + \sigma^2)} \\ &= c \sqrt{\frac{d}{n} ((\|\theta_1^*\|^2 + \|\theta_2^*\|^2)/2 + \sigma^2) \log(1/\delta)}, \end{aligned}$$

for some  $c > 0$  with probability at least  $1 - \delta$ . □

**Lemma 3.** *There exists constant  $D_2 > 0$  for which*

$$\mathbb{P} \left\{ \|M_n(\theta, \underline{X}, \underline{Y}^{(s)}) - M_n(\theta, \underline{X}, \tilde{\underline{Y}})\| \leq D_2 \|\hat{s} - s\| \right\} \rightarrow 1$$

for all  $\theta \in \mathbb{R}^d$ .

*Proof.* Write

$$\begin{aligned} &M_n(\theta, \underline{X}, \underline{Y}^{(s)}) - M_n(\theta, \underline{X}, \tilde{\underline{Y}}) \\ &= \hat{\Sigma}^{-1} \frac{2}{n} \sum_{i=1}^n \left[ \phi(Y_i^{(s)} \langle \theta, X_i \rangle) X_i Y_i^{(s)} - \phi(\tilde{Y}_i \langle \theta, X_i \rangle) X_i \tilde{Y}_i \right] + \hat{\Sigma}^{-1} \frac{1}{n} \sum_{i=1}^n X_i (\tilde{Y}_i - Y_i^{(s)}). \end{aligned}$$

Use triangle inequality to deduce that

$$\begin{aligned} \|M_n(\theta, \underline{X}, \underline{Y}^{(s)}) - M_n(\theta, \underline{X}, \tilde{\underline{Y}})\| &\leq \left\| \hat{\Sigma}^{-1} \frac{2}{n} \sum_{i=1}^n \phi(Y_i^{(s)} \langle \theta, X_i \rangle) X_i \langle \hat{s} - s, X_i \rangle \right\| \\ &\quad + \left\| \hat{\Sigma}^{-1} \frac{2}{n} \sum_{i=1}^n \left( \phi(Y_i^{(s)} \langle \theta, X_i \rangle) - \phi(\tilde{Y}_i \langle \theta, X_i \rangle) \right) X_i \tilde{Y}_i \right\| \\ &\quad + \left\| \hat{\Sigma}^{-1} \frac{1}{n} \sum_{i=1}^n X_i X_i^T (\hat{s} - s) \right\|. \end{aligned}$$

The first and the third term can be bounded by a constant multiple of  $\|\hat{s} - s\|$  with high probability. Simply notice that  $\mathbb{P}\{\|\hat{\Sigma}^{-1}\|_{op} > 2\} \rightarrow 0$  and  $|\phi| \leq 1$ . For the second term, use the mean-value theorem and the basic inequality  $|u\phi'(u)| < e^{-|u|}$  for all  $u \in \mathbb{R}$  to bound this term by

$$\|\hat{\Sigma}^{-1}\|_{op} \frac{2}{n} \sum_{i=1}^n \left| \frac{\tilde{Y}_i}{Y_i^{(m)}} \right| \exp(-|Y_i^{(m)} \langle \theta, X_i \rangle|) \cdot \|X_i\| \|\hat{s} - s\|$$

for some  $Y_i^{(m)}$  that lies between  $Y_i^{(s)}$  and  $\tilde{Y}_i$ . The above is bounded by a constant multiple of  $\|\hat{s} - s\|$  with high probability. □

**Theorem 3.** Apply the sample-splitting version of EM discribed in Section 3 on the shifted data  $\tilde{Y}$  defined above and assume that  $\theta_0$  satisfies the same initialization conditions with  $\theta^* = (\theta_1^* - \theta_2^*)/2$ . There exists constant  $C > 0$  for which the EM iterates  $\{\theta^t\}_{t=1}^T$  satisfy

$$\begin{aligned}\|\theta^t - \theta^*\| &\leq \gamma^t \|\theta^0 - \theta^*\| + \frac{C \sqrt{\|\theta^*\|^2 + \|s\|^2 + \sigma^2}}{1 - \gamma} \sqrt{\frac{dT \log(T/\delta)}{n}} \\ &= \gamma^t \|\theta^0 - \theta^*\| + \frac{C \sqrt{(\|\theta_1^*\|^2 + \|\theta_2^*\|^2)/2 + \sigma^2}}{1 - \gamma} \sqrt{\frac{dT \log(T/\delta)}{n}}.\end{aligned}$$

with probability at least  $1 - \delta$ .

*Proof.* Follow the proof of Theorem 2 which gives convergence rates for the symmetric EM iterates. We have

$$\|\theta^{t+1} - \theta^*\| \leq \gamma \|\theta^t - \theta^*\| + \max_{t \in [T]} \|M_{n/T}(\theta^t, \underline{X}, \tilde{Y}) - M(\theta^t)\| + \max_{t \in [T]} \|M_{n/T}(\theta^t, \underline{X}, \tilde{Y}) - M_{n/T}(\theta^t, \underline{X}, Y^{(s)})\|.$$

The second term was handled in the proof of Theorem 2. We only need to bound the third term. It suffices to show that

$$\|M_{n/T}(\theta^t, \underline{X}, \tilde{Y}) - M_{n/T}(\theta^t, \underline{X}, Y^{(s)})\| \leq \epsilon_S(n/T, \delta/T)$$

with probability at least  $1 - \delta/T$ . We need

$$\epsilon_S(n, \delta) \leq D_3 \sqrt{\frac{d}{n} (\|s\|^2 + \|\theta^*\|^2 + \sigma^2) \log \frac{1}{\delta}}$$

for some  $D_3 > 0$ . That is an easy consequence of Lemma 2 and Lemma 3. The rest of the proof follows exactly as that of Theorem 2.  $\square$

**Remark 3.** Combine Lemma 2 and Theorem 3 to deduce the error rates on the original centers.

$$\|\theta_i^t - \theta_i^*\| \leq \gamma^t \|\theta^0 - \theta^*\| + \frac{(C + D_1) \sqrt{(\|\theta_1^*\|^2 + \|\theta_2^*\|^2)/2 + \sigma^2}}{1 - \gamma} \sqrt{\frac{dT \log(T/\delta)}{n}},$$

for  $i = 1, 2$  with probability at least  $1 - \delta$ .

## 5 Proofs of main theorems

*Proof of Lemma 1.* For simplification, we assume throughout this proof that  $\sigma^2 = 1$ . If  $W = \langle \theta^*, X \rangle + \varepsilon$ , a few applications of Stein's Lemma [12, Lemma 1] yields

$$\begin{aligned}M(\theta) &= \mathbb{E}[(2\phi(W\langle \theta, X \rangle) - 1)XW] \\ &= \theta^* (\mathbb{E}[2\phi(W\langle \theta, X \rangle) + 2W\langle \theta, X \rangle \phi'(W\langle \theta, X \rangle) - 1]) + \theta \mathbb{E}[2W^2 \phi'(W\langle \theta, X \rangle)].\end{aligned}$$

In what follows, we let

$$A = \mathbb{E}[2\phi(W\langle \theta, X \rangle) + 2W\langle \theta, X \rangle \phi'(W\langle \theta, X \rangle) - 1]$$

and

$$B = 2W^2 \phi'(W\langle \theta, X \rangle).$$

Thus, we see that  $M(\theta) = \theta^*A + \theta B$  belongs to  $\text{span}\{\theta, \theta^*\} = \{\lambda_1\theta + \lambda_2\theta^*, : \lambda_1, \lambda_2 \in \mathbb{R}\}$ . This is a crucial fact that will exploit multiple times.

Observe that for any  $a$  in  $\text{span}\{\theta, \theta^*\}$ ,

$$a = \langle \theta_0, a \rangle \theta_0 + \langle \theta_0^\perp, a \rangle \theta_0^\perp,$$

and

$$\|a\|^2 = |\langle \theta_0, a \rangle|^2 + |\langle \theta_0^\perp, a \rangle|^2.$$

Specializing this to  $a = M(\theta) - \theta^*$  yields

$$\|M(\theta) - \theta^*\|^2 = |\langle \theta_0, M(\theta) - \theta^* \rangle|^2 + |\langle \theta_0^\perp, M(\theta) - \theta^* \rangle|^2.$$

The strategy for establishing contractivity of  $M(\theta)$  will be to show that the sum of  $|\langle \theta_0, M(\theta) - \theta^* \rangle|^2$  and  $|\langle \theta_0^\perp, M(\theta) - \theta^* \rangle|^2$  is less than  $\gamma^2 \|\theta - \theta^*\|^2$ . This idea was used in [5] to obtain global contractivity of the population EM operator for the mixture of two Gaussians problem.

To reduce this  $(d+1)$ -dimensional problem (as seen from the joint distribution of  $(X, Y)$ ) to a 2-dimensional problem, we note that

$$W\langle \theta, X \rangle \stackrel{\mathcal{D}}{=} \Lambda Z_1 Z_2 + \Gamma Z_2^2,$$

where  $Z_1, Z_2 \stackrel{i.i.d.}{\sim} N(0, 1)$ . The coefficients  $\Gamma$  and  $\Lambda$  are

$$\Gamma = \langle \theta, \theta^* \rangle$$

and

$$\Lambda^2 = \|\theta\|^2(1 + \|\theta^*\|^2) - \Gamma^2 = \|\theta\|^2(1 + |\langle \theta_0^\perp, \theta^* \rangle|^2).$$

This is because we have

$$(W, \langle \theta, X \rangle) \stackrel{\mathcal{D}}{=} (\sqrt{1 + \|\theta^*\|^2} Z_2, \frac{\Lambda}{\sqrt{1 + \|\theta^*\|^2}} Z_1 + \frac{\Gamma}{\sqrt{1 + \|\theta^*\|^2}} Z_2).$$

Note that  $\Lambda Z_1 Z_2 + \Gamma Z_2^2 \stackrel{\mathcal{D}}{=} \Lambda Z_1 |Z_2| + \Gamma Z_2^2$  because they have the same moment generating function. Deduce that

$$W\langle \theta, X \rangle \stackrel{\mathcal{D}}{=} \Lambda Z_1 |Z_2| + \Gamma Z_2^2.$$

Lemma 9 implies that

$$(1 - \kappa) \langle \theta_0^\perp, \theta^* \rangle \leq \langle \theta_0^\perp, M(\theta) \rangle \leq (1 + \sqrt{\kappa}) \langle \theta_0^\perp, \theta^* \rangle,$$

and consequently,

$$|\langle \theta_0^\perp, M(\theta) - \theta^* \rangle| \leq \sqrt{\kappa} |\langle \theta_0^\perp, \theta - \theta^* \rangle| \leq \sqrt{\kappa} \|\theta - \theta^*\|. \quad (7)$$

Next, we note that  $\Lambda^2 \rightarrow \Gamma$  as  $\theta \rightarrow \theta^*$ . In fact,

$$\begin{aligned} |\Lambda^2 - \Gamma| &= |\|\theta\|^2(1 + |\langle \theta_0^\perp, \theta^* \rangle|^2) - \langle \theta, \theta^* \rangle| \\ &\leq \|\theta\|^2 |\langle \theta_0^\perp, \theta^* \rangle|^2 + |\langle \theta, \theta - \theta^* \rangle| \\ &\leq \|\theta\| (|\langle \theta_0^\perp, \theta^* \rangle| + 1) \|\theta - \theta^*\|. \end{aligned}$$

Next, define

$$h(\alpha, \beta) = \mathbb{E}[(2\phi(\alpha Z_2(Z_1 + \beta Z_2)) - 1)(Z_2(Z_1 + \beta Z_2))].$$



Note that by definition of  $h$  and Lemma 6,  $h(\Lambda, \frac{\Gamma}{\Lambda}) = \frac{\langle \theta, M(\theta) \rangle}{\Lambda}$ . In fact,  $h$  is the one-dimensional population EM operator for this model. By the self-consistency property of EM [10, page 79],  $h(\beta, \beta) = \beta$ . Translating this to our problem, we have that  $h(\frac{\Gamma}{\Lambda}, \frac{\Gamma}{\Lambda}) = \frac{\Gamma}{\Lambda} = \frac{\langle \theta, \theta^* \rangle}{\Lambda}$ . Since  $h(\Lambda, \frac{\Gamma}{\Lambda}) - h(\frac{\Gamma}{\Lambda}, \frac{\Gamma}{\Lambda}) = \int_{\frac{\Gamma}{\Lambda}}^{\Lambda} \frac{\partial h}{\partial \alpha} h(\alpha, \frac{\Gamma}{\Lambda}) d\alpha$ , we have from Lemma 10,

$$\begin{aligned} |\langle \theta_0, M(\theta) - \theta^* \rangle| &\leq \frac{\Lambda}{\|\theta\|} \left| \int_{\frac{\Gamma}{\Lambda}}^{\Lambda} \frac{\partial h}{\partial \alpha} h(\alpha, \frac{\Gamma}{\Lambda}) d\alpha \right| \\ &\leq \frac{2\Lambda}{\|\theta\|} \sqrt{\kappa} \left| \int_{\frac{\Gamma}{\Lambda}}^{\Lambda} \frac{d\alpha}{\alpha^2} \right| \\ &= 2\sqrt{\kappa} \frac{|\Lambda^2 - \Gamma|}{\Gamma \|\theta\|} \\ &\leq 2\sqrt{\kappa} \left( \frac{|\langle \theta^\perp, \theta^* \rangle| + 1}{\langle \theta, \theta^* \rangle} \right) \|\theta - \theta^*\|. \end{aligned}$$

Combining this with inequality (7) yields (3).  $\square$

**Remark 4.** The function  $h$  is related to the EM operator for the one-dimensional symmetric mixture of two Gaussians model

$$Y = R\beta + \varepsilon,$$

$R \sim \text{Rademacher}(1/2)$  and  $\varepsilon \sim N(0, 1)$ . One can derive that (see [9, page 4]) the population EM operator is

$$T(\alpha, \beta) = \mathbb{E}[(2\phi(\alpha(Z_1 + \beta)) - 1)(Z_1 + \beta)].$$

Then  $h(\alpha, \beta)$  is a “smoothed” version of  $T(\alpha, \beta)$  as seen through the identity

$$h(\alpha, \beta) = \mathbb{E}[|Z_2|T(\alpha|Z_2|, \beta|Z_2|)].$$

In light of this relationship, it is perhaps not surprising that the EM operator for the mixture of linear regressions problem also enjoys a large basin of attraction.

**Remark 5.** Recently in [2], the authors analyzed gradient descent for a single-hidden layer convolutional neural network structure with no overlap and Gaussian input. In this setup, we observe i.i.d. data  $(X_i, Y_i)_{i=1}^n$ , where  $Y_i = f(X_i, w) + \varepsilon_i$  and  $X_i \sim N(0, I_d)$  and  $\varepsilon_i \sim N(0, \sigma^2)$  are independent of each other. The neural network has the form  $f(x, w) = \frac{1}{k} \sum_{j=1}^k \max\{0, \langle w_j, x \rangle\}$  and the only nonzero coordinates of  $w_j$  are in the  $j$ -th successive block of  $d/k$  coordinates and are equal to a fixed  $d/k$  dimensional filter vector  $w$ . One desires to minimize the risk  $\ell(w) = \mathbb{E}(f(X, w) - f(X, w^*))^2$ . Interestingly, the gradient of  $\ell(w)$  belongs to the linear span of  $w$  and  $w^*$ , akin to our  $M(\theta) \in \text{span}\{\theta, \theta^*\}$  (and also in the Gaussian mixture problem [9]). This property plays a critical role in the analysis.

One can use an alternative scheme to gradient descent using a simple method of moments estimator based on the identity  $2\mathbb{E}[X \max\{0, \langle w, X \rangle\}] = w$ . We observe that  $\hat{w} = \frac{2}{n} \sum_{i=1}^n X_i Y_i$  is an unbiased estimator of  $\frac{1}{k} \sum_{j=1}^k w_j^*$  (in fact,  $w^*$  need not be the same across successive blocks) and its mean square error is less than a multiple of  $\frac{d}{n}(\|w^*\|^2 + \sigma^2) \log(1/\delta)$  with probability at least  $1 - \delta$ . Our problem, however, is not directly amenable to such a method.

*Proof of Theorem 1.* Note that in general,  $M(\theta) = \theta^* A + \theta B$ , where

$$A = \mathbb{E}[2\phi(W\langle \theta, X \rangle / \sigma^2) + 2(W\langle \theta, X \rangle / \sigma^2)\phi'(W\langle \theta, X \rangle / \sigma^2) - 1],$$

$$B = 2\mathbb{E}[(W^2/\sigma^2)\phi'(W\langle\theta, X\rangle/\sigma^2)].$$

The assumption  $\langle\theta, \theta^\star\rangle = 0$  implies that  $A = 0$ . To see this, note that

$$\mathbb{E}\phi(W\langle\theta, X\rangle) = \mathbb{E}\phi(\Lambda Z_1 Z_2) = \phi(0) = 1/2,$$

and

$$\mathbb{E}[W\langle\theta, X\rangle\phi'(W\langle\theta, X\rangle)] = \mathbb{E}[\Lambda Z_1 Z_2\phi'(\Lambda Z_1 Z_2)] = 0,$$

Finally, observe that  $B = 2(1 + \|\theta^\star\|^2/\sigma^2)\mathbb{E}[Z_2^2\phi'(Z_1 Z_2\|\theta\|\sqrt{\sigma^2 + \|\theta^\star\|^2}/\sigma^2)] \rightarrow 1 + \|\theta^\star\|^2/\sigma^2 > 1$  as  $\theta \rightarrow 0$ . Thus, there are choices of  $\theta$  for which  $B > 1$  and hence

$$\begin{aligned}\|M(\theta) - \theta^\star\|^2 &= \|\theta - \theta^\star\|^2 + (B^2 - 1)\|\theta\|^2 \\ &> \|\theta - \theta^\star\|^2.\end{aligned}$$

□

*Proof of Theorem 2.* The conditions on  $\rho$ ,  $\|\theta\|$ , and  $\|\theta^\star\|$  ensure that the factor on the right side of inequality (3) multiplying  $\|\theta - \theta^\star\|$  is less than  $1/4$ .

Consider the set  $\mathcal{A} = \{\theta : \langle\theta, \theta^\star\rangle > \rho\|\theta\|\|\theta^\star\|, 3\sigma \leq \|\theta\| \leq R\sigma\}$ . We will show that the empirical EM updates stay in this set. That is,  $M_n(\mathcal{A}) \subseteq \mathcal{A}$ . This is based on Lemma 4 which shows that

$$M(\mathcal{A}) \subseteq \{\theta : \langle\theta, \theta^\star\rangle > (1 + \Delta)\rho\|\theta\|\|\theta^\star\|, \|\theta^\star\|(1 - \kappa) \leq \|\theta\| \leq \sqrt{\sigma^2 + 3\|\theta^\star\|^2}\}.$$

This statement is what allows us to say that  $M_n(\mathcal{A}) \subseteq \mathcal{A}$ ; in particular when  $M_n$  is close to  $M$ . To be precise, assume  $\sup_{\theta \in \mathcal{A}} \|M_n(\theta) - M(\theta)\| < \epsilon$ . That implies

$$\sup_{\theta \in \mathcal{A}} \left\| \frac{M_n(\theta)}{\|M_n(\theta)\|} - \frac{M(\theta)}{\|M(\theta)\|} \right\| \leq \sup_{\theta \in \mathcal{A}} \frac{2\|M_n(\theta) - M(\theta)\|}{\|M(\theta)\|} < \frac{2\epsilon}{(1 - \kappa)\|\theta^\star\|}.$$

For the last inequality, we used the fact that  $\|M(\theta)\| \geq \|\theta^\star\|(1 - \kappa)$  for all  $\theta$  in  $\mathcal{A}$ . It follows from Lemma 4 that

$$\begin{aligned}\sup_{\theta \in \mathcal{A}} \langle\theta^\star, \frac{M_n(\theta)}{\|M_n(\theta)\|}\rangle &\geq \sup_{\theta \in \mathcal{A}} \langle\theta^\star, \frac{M(\theta)}{\|M(\theta)\|}\rangle - \frac{2\epsilon}{(1 - \kappa)} \\ &\geq \|\theta^\star\|(1 + \Delta)\rho - \frac{2\epsilon}{(1 - \kappa)} \\ &\geq \|\theta^\star\|\rho,\end{aligned}$$

provided  $\epsilon < (\frac{1 - \kappa}{2})\Delta\rho\|\theta^\star\|$  and

$$\begin{aligned}\sup_{\theta \in \mathcal{A}} \|M_n(\theta)\| &\geq \sup_{\theta \in \mathcal{A}} \|M(\theta)\| - \epsilon \\ &\geq \|\theta^\star\|(1 - \kappa) - \epsilon \\ &\geq 4\sigma(1 - \kappa) - \epsilon \\ &\geq 3\sigma,\end{aligned}$$

provided  $\epsilon < \sigma(1 - 4\kappa)$ . Also, note that

$$\begin{aligned}\sup_{\theta \in \mathcal{A}} \|M_n(\theta)\| &\leq \sup_{\theta \in \mathcal{A}} \|M(\theta)\| + \epsilon \\ &\leq \sqrt{\sigma^2 + 3\|\theta^\star\|^2} + \epsilon \\ &\leq R\sigma,\end{aligned}$$

provided  $\epsilon < R\sigma - \sqrt{\sigma^2 + 3\|\theta^*\|^2}$ . For this to be true, we also require that  $R$  be large enough so that  $R\sigma - \sqrt{\sigma^2 + 3\|\theta^*\|^2} > 0$ .

For  $\delta \in (0, 1)$ , let  $\epsilon_M(n, \delta)$  be the smallest number such that for any fixed  $\theta$  in  $\mathcal{A}$ , we have

$$\|M_n(\theta) - M(\theta)\| \leq \epsilon_M(n, \delta),$$

with probability at least  $1 - \delta$ . Moreover, suppose  $n$  is large enough so that

$$\epsilon_M(n, \delta) \leq \min\{\sigma(1 - 4\kappa), (\frac{1-\kappa}{2})\Delta\rho\|\theta^*\|, R\sigma - \sqrt{\sigma^2 + 3\|\theta^*\|^2}\},$$

which guarantees that  $M_n(\mathcal{A}) \subseteq \mathcal{A}$ . For any iteration  $t \in [T]$ , we have

$$\|M_{n/T}(\theta^t) - M(\theta^t)\| \leq \epsilon_M(n/T, \delta/T),$$

with probability at least  $1 - \delta/T$ . Thus by a union bound and  $M_n(\mathcal{A}) \subseteq \mathcal{A}$ ,

$$\max_{t \in [T]} \|M_{n/T}(\theta^t) - M(\theta^t)\| \leq \epsilon_M(n/T, \delta/T),$$

with probability at least  $1 - \delta$ .

Hence if  $\theta^0$  belongs to  $\mathcal{A}$ , then by Lemma 1,

$$\begin{aligned} \|\theta^{t+1} - \theta^*\| &= \|M_{n/T}(\theta^t) - \theta^*\| \\ &\leq \|M(\theta^t) - \theta^*\| + \|M_{n/T}(\theta^t) - M(\theta^t)\| \\ &\leq \gamma\|\theta^t - \theta^*\| + \max_{t \in [T]} \|M_{n/T}(\theta) - M(\theta)\| \\ &\leq \gamma\|\theta^t - \theta^*\| + \epsilon_M(n/T, \delta/T). \end{aligned}$$

Solving this recursive inequality yields,

$$\begin{aligned} \|\theta^t - \theta^*\| &\leq \gamma^t\|\theta^0 - \theta^*\| + \epsilon_M(n/T, \delta/T) \sum_{i=0}^{t-1} \gamma^i \\ &\leq \gamma^t\|\theta^0 - \theta^*\| + \frac{\epsilon_M(n/T, \delta/T)}{1-\gamma}, \end{aligned}$$

with probability at least  $1 - \delta$ .

Finally, it was shown in [1] that

$$\epsilon_M(n/T, \delta/T) \leq C\sqrt{\|\theta^*\|^2 + \sigma^2} \sqrt{\frac{dT \log(T/\delta)}{n}}$$

with probability at least  $1 - \delta/T$ .

□

## 6 Discussion

In this paper, we showed that the empirical EM iterates converge to true coefficients of a mixture of two linear regressions as long as the initializer lies within a “cone” or “wedge” shape region (see the condition on Theorem 2:  $\langle \theta^0, \theta^* \rangle > \rho\|\theta^0\|\|\theta^*\|$ ).

In Fig. 2, the red points symmetric around zero are  $\theta^*$  and  $-\theta^*$ . All entries of the design matrix  $X$  and the noise  $\varepsilon$  is generated *i.i.d.* from the standard normal distribution. We took 100 initializers generated uniformly from the  $[-5, 5] \times [-5, 5]$  square, ran the EM algorithm with  $T = 100$  and plotted the  $T$ -th EM iterate for each initializer.

As seen from Fig. 2, the EM updates lead to the truth regardless of the location of the initializer. However, Theorem 1 implies that this is not always the case, since convergence can fail if  $\theta^0$  very small and  $\langle \theta^0, \theta^* \rangle = 0$ .

Finally, we provide a plot (Fig. 3) depicting the traces of the EM iterates. It appears that the EM algorithm tends to pull the initializers onto an ellipse in very few steps. The iterates then travel along this ellipse, in the direction of the truth.

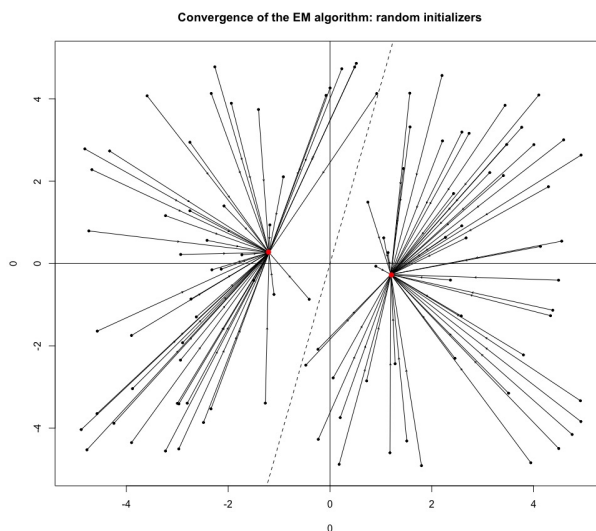


Figure 2: Convergence of the empirical EM iterates with random initialization.

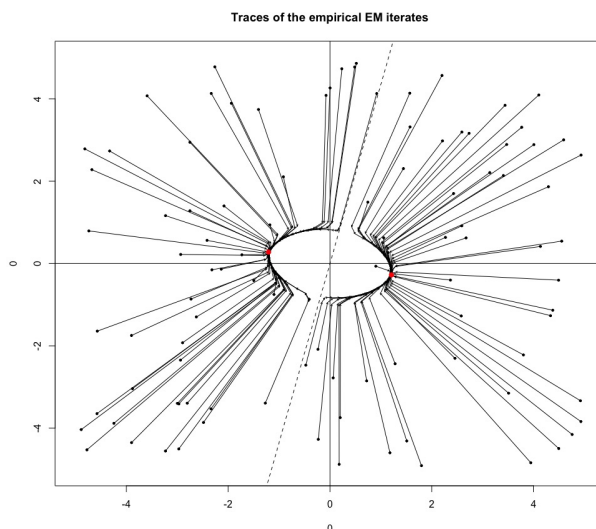


Figure 3: Traces of the empirical EM iterates.

## A Appendix

For the following lemma, let

$$A = \mathbb{E}[2\phi(W\langle\theta, X\rangle/\sigma^2) + 2(W\langle\theta, X\rangle/\sigma^2)\phi'(W\langle\theta, X\rangle/\sigma^2) - 1],$$

$$B = 2\mathbb{E}[(W^2/\sigma^2)\phi'(W\langle\theta, X\rangle/\sigma^2)],$$

and

$$\kappa = \frac{1}{\sqrt{\frac{\Gamma}{\Lambda} \min\{\Lambda, \frac{\Gamma}{\Lambda}\} + 1}} = \max \left\{ \left(1 - \frac{|\langle\theta_0, \theta^*\rangle|^2}{\sigma^2 + \|\theta^*\|^2}\right)^{1/2}, \left(\frac{\sigma^2}{\sigma^2 + \langle\theta, \theta^*\rangle}\right)^{1/2} \right\}.$$

**Lemma 4.** *The cosine angle between  $\theta^*$  and  $M(\theta)$  is equal to*

$$\frac{\|\theta^*\|^2 A + \langle\theta, \theta^*\rangle B}{\sqrt{(\|\theta^*\|^2 A + \langle\theta, \theta^*\rangle B)^2 + B^2(\|\theta\|^2 \|\theta^*\|^2 - |\langle\theta, \theta^*\rangle|^2)}}.$$

If  $\langle\theta, \theta^*\rangle \geq \rho\|\theta\|\|\theta^*\|$  and  $3\sigma \leq \|\theta\| \leq R\sigma$ , then there exists positive  $\Delta = \Delta(\rho, \sigma, \|\theta^*\|, R)$  such that this cosine angle is at least  $(1 + \Delta)\rho$ . Moreover,

$$\|\theta^*\|^2(1 - \kappa)^2 \leq \|M(\theta)\|^2 = \|\theta^*\|^2 A^2 + \|\theta\|^2 B^2 + 2\langle\theta, \theta^*\rangle AB \leq \sigma^2 + 3\|\theta^*\|^2,$$

and

$$\langle\theta^*, M(\theta)\rangle = \|\theta^*\|^2 A + \langle\theta, \theta^*\rangle B \geq \|\theta^*\|^2(1 - \kappa).$$

*Proof.* We will prove the first statement. Let  $\tau = \frac{\|\theta^*\|}{\|\theta\|} \frac{A}{B}$ . Observe that

$$\begin{aligned} \frac{\|\theta^*\|^2 A + \langle\theta, \theta^*\rangle B}{\sqrt{(\|\theta^*\|^2 A + \langle\theta, \theta^*\rangle B)^2 + B^2(\|\theta\|^2 \|\theta^*\|^2 - |\langle\theta, \theta^*\rangle|^2)}} &= \frac{1}{\sqrt{1 + \frac{\|\theta\|^2 \|\theta^*\|^2 - |\langle\theta, \theta^*\rangle|^2}{(\|\theta^*\|^2 \frac{A}{B} + \langle\theta, \theta^*\rangle)^2}}} \\ &\geq \frac{1}{\sqrt{1 + \frac{1 - \rho^2}{(\tau + \rho)^2}}} \\ &= \frac{\rho}{\sqrt{1 - (1 - \rho^2) \frac{\tau(\tau + 2\rho)}{(\tau + \rho)^2}}} \\ &\geq \frac{\rho}{\sqrt{1 - (1 - \rho^2) \frac{\tau}{\tau + \rho}}} \\ &\geq \rho(1 + \frac{1}{2}(1 - \rho^2) \frac{\tau}{\tau + \rho}), \end{aligned}$$

where the last line follows from the inequality  $1/\sqrt{1 - a} \geq 1 + a/2$  for all  $a \in (0, 1)$ .

Finally, note that from Lemma 9,

$$\frac{A}{B} \geq \frac{\sigma^2(1 - \kappa)}{2(\|\theta^*\|^2 + \sigma^2)\kappa^3}.$$

Thus,  $\tau \geq \tau_0 := \frac{\sigma\|\theta^*\|(1 - \kappa)}{2R(\|\theta^*\|^2 + \sigma^2)\kappa^3}$  and so we can set

$$\Delta = \frac{1}{2}(1 - \rho^2) \frac{\tau_0}{\tau_0 + \rho} > 0.$$

For the second claim, the identity

$$\|M(\theta)\|^2 = \|\theta^*\|^2 A^2 + \|\theta\|^2 B^2 + 2\langle\theta, \theta^*\rangle AB$$

is an immediate consequence of  $M(\theta) = A\theta^* + B\theta$ . By Lemma 9,  $A \geq 1 - \kappa$  and hence since  $\langle \theta, \theta^* \rangle \geq 0$ , we have  $\|M(\theta)\|^2 \geq \|\theta^*\|^2 A^2 \geq \|\theta^*\|^2 (1 - \kappa)^2$ .

Next, we will show that  $\|M(\theta)\|^2 \leq \sigma^2 + 3\|\theta^*\|^2$ . To see this, note that by Lemma 6 and Jensen's inequality,

$$\begin{aligned} \langle \theta, M(\theta) \rangle &= \mathbb{E}[(2\phi(W\langle \theta, X \rangle) - 1)W\langle \theta, X \rangle] \\ &\leq \mathbb{E}|W\langle \theta, X \rangle| \\ &\leq \sqrt{\mathbb{E}|W\langle \theta, X \rangle|^2} \\ &= \sqrt{\Lambda^2 + 3\Gamma^2} \\ &= \|\theta\| \sqrt{\sigma^2 + \|\theta^*\|^2 + 2|\langle \theta_0, \theta^* \rangle|^2}. \end{aligned}$$

Next, it can be shown that  $A \leq \sqrt{2}$  and hence

$$\begin{aligned} \langle \theta_0^\perp, M(\theta) \rangle &= A\langle \theta_0^\perp, \theta^* \rangle \\ &\leq \sqrt{2}\langle \theta_0^\perp, \theta^* \rangle. \end{aligned}$$

Putting these two facts together, we have

$$\begin{aligned} \|M(\theta)\|^2 &= |\langle \theta_0^\perp, M(\theta) \rangle|^2 + |\langle \theta_0, M(\theta) \rangle|^2 \\ &\leq \sigma^2 + \|\theta^*\|^2 + 2|\langle \theta_0^\perp, \theta^* \rangle|^2 + 2|\langle \theta_0, \theta^* \rangle|^2 \\ &= \sigma^2 + 3\|\theta^*\|^2. \end{aligned}$$

The final statement

$$\langle \theta^*, M(\theta) \rangle = \|\theta^*\|^2 A + \langle \theta, \theta^* \rangle B \geq \|\theta^*\|^2 (1 - \kappa).$$

follows from similar arguments. □

**Lemma 5.** *If  $\langle \theta, \theta^* \rangle \geq 0$  and  $\sigma^2 = 1$ , then*

$$\mathbb{E}[W\langle \theta, X \rangle \phi'(W\langle \theta, X \rangle)] \geq 0.$$

*Proof.* Note that the statement is true if

$$\mathbb{E}[(\alpha Z + \beta)\phi'(\alpha Z + \beta)] \geq 0,$$

where  $Z \sim N(0, 1)$  and  $\alpha \geq 0$  and  $\beta \geq 0$ . This fact is proved in Lemma 5 in [9] or Lemma 1 in [5]. □

**Lemma 6.** *Assume  $\sigma^2 = 1$ . Then*

$$\begin{aligned} \langle \theta, M(\theta) \rangle &= \mathbb{E}[(2\phi(W\langle \theta, X \rangle) - 1)W\langle \theta, X \rangle] \\ &= \mathbb{E}[(2\phi(\Lambda Z_1 Z_2 + \Gamma Z_2^2) - 1)(\Lambda Z_1 Z_2 + \Gamma Z_2^2)], \end{aligned}$$

and

$$\langle \theta_0^\perp, M(\theta) \rangle = \langle \theta_0^\perp, \theta^* \rangle \mathbb{E}[2\phi(W\langle \theta, X \rangle) + 2W\langle \theta, X \rangle \phi'(W\langle \theta, X \rangle) - 1].$$

**Lemma 7.** *The following inequalities hold for all  $x \in \mathbb{R}$ :*

$$|2\phi(x) + 2x\phi'(x) - 1| \leq 1 + \sqrt{2(1 - \phi(x))},$$

and

$$x^2\phi'(x) \leq \sqrt{2(1 - \phi(x))}.$$

**Lemma 8.** Let  $\alpha, \beta > 0$  and  $Z \sim N(0, 1)$ . Then

$$\mathbb{E}2(1 - \phi(\alpha(Z + \beta))) \leq \exp\{-\frac{\beta}{2} \min\{\alpha, \beta\}\}.$$

Moreover,

$$\mathbb{E}2(1 - \phi(\alpha Z_2(Z_1 + \beta Z_2))) \leq \frac{1}{\sqrt{\beta \min\{\alpha, \beta\} + 1}}.$$

*Proof.* The second conclusion follows immediately from the first since

$$\begin{aligned} \mathbb{E}2(1 - \phi(\alpha Z_2(Z_1 + \beta Z_2))) &= \mathbb{E}_{Z_2} \mathbb{E}_{Z_1} 2(1 - \phi(\alpha |Z_2| (Z_1 + \beta |Z_2|))) \\ &\leq \mathbb{E}_{Z_2} \exp\{-\frac{Z_2^2}{2} \beta \min\{\alpha, \beta\}\} \\ &= \frac{1}{\sqrt{\beta \min\{\alpha, \beta\} + 1}}. \end{aligned}$$

The last equality follows from the moment generating function of  $\chi^2(1)$ .

For the first conclusion, we first observe that the mapping  $\alpha \mapsto \mathbb{E}\phi(\alpha(Z + \beta))$  is increasing (Lemma 5 in [9] or Lemma 1 in [5]). Next, note the inequality

$$2(1 - \phi(x)) \leq e^{-x},$$

which is equivalent to  $(e^x - 1)^2 \geq 0$ . If  $\alpha \geq \beta$ , then

$$\begin{aligned} \mathbb{E}2(1 - \phi(\alpha(Z + \beta))) &\leq \mathbb{E}2(1 - \phi(\beta(Z + \beta))) \\ &\leq \mathbb{E}e^{-(\beta(Z + \beta))} \\ &= e^{-\beta^2/2}. \end{aligned}$$

If  $\alpha \leq \beta$ , then

$$\begin{aligned} \mathbb{E}2(1 - \phi(\alpha(Z + \beta))) &\leq \mathbb{E}e^{-(\alpha(Z + \beta))} \\ &= e^{\alpha^2/2 - \alpha\beta} \\ &\leq e^{-\alpha\beta/2}. \end{aligned}$$

In each case, we used the moment generating function of a normal distribution to evaluate the expectations.  $\square$

**Lemma 9.** Assume  $\sigma^2 = 1$ . We have

$$1 - \kappa \leq A \leq 1 + \sqrt{\kappa},$$

and

$$B \leq 2(1 + \|\theta^*\|^2)\kappa^3.$$

*Proof.* By Lemma 5 and Lemma 8,

$$\begin{aligned} A &= \mathbb{E}[2\phi(W\langle\theta, X\rangle) + 2W\langle\theta, X\rangle\phi'(W\langle\theta, X\rangle) - 1] \\ &\geq \mathbb{E}[2\phi(W\langle\theta, X\rangle) - 1] \\ &\geq 1 - \kappa. \end{aligned}$$

By Lemma 7, Jensen's inequality, and Lemma 8,

$$\begin{aligned}
A &= \mathbb{E}[2\phi(W\langle\theta, X\rangle) + 2W\langle\theta, X\rangle\phi'(W\langle\theta, X\rangle) - 1] \\
&\leq \mathbb{E}[1 + \sqrt{2(1 - \phi(W\langle\theta, X\rangle))}] \\
&\leq 1 + \sqrt{\mathbb{E}2(1 - \phi(W\langle\theta, X\rangle))} \\
&\leq 1 + \sqrt{\kappa}.
\end{aligned}$$

By Lemma 8,

$$\begin{aligned}
B &= 2\mathbb{E}[W^2\phi'(W\langle\theta, X\rangle)] \\
&\leq 2\mathbb{E}[2W^2(1 - \phi(W\langle\theta, X\rangle))] \\
&= 2(1 + \|\theta^*\|^2)\mathbb{E}_{Z_2}Z_2^2\mathbb{E}_{Z_1}[2(1 - \phi(\Lambda Z_2(Z_1 + \frac{\Gamma}{\Lambda}Z_2)))] \\
&\leq 2(1 + \|\theta^*\|^2)\mathbb{E}_{Z_2}[Z_2^2 \exp\{-\frac{Z_2^2}{2}\frac{\Gamma}{\Lambda} \min\{\frac{\Gamma}{\Lambda}, \Lambda\}\}] \\
&= 2(1 + \|\theta^*\|^2)(\frac{1}{\frac{\Gamma}{\Lambda} \min\{\Lambda, \frac{\Gamma}{\Lambda}\} + 1})^{3/2} \\
&= 2(1 + \|\theta^*\|^2)\kappa^3.
\end{aligned}$$

□

**Lemma 10.** *Define*

$$h(\alpha, \beta) = \mathbb{E}[(2\phi(\alpha Z_2(Z_1 + \beta Z_2)) - 1)(Z_2(Z_1 + \beta Z_2))].$$

Let  $\alpha, \beta > 0$ . Then

$$\frac{\partial}{\partial \alpha} h(\alpha, \beta) \leq \frac{2}{\alpha^2} (\frac{1}{\beta \min\{\alpha, \beta\} + 1})^{1/4}.$$

*Proof.* First, observe that

$$\frac{\partial}{\partial \alpha} h(\alpha, \beta) = \mathbb{E}[2\phi'(\alpha Z_2(Z_1 + \beta Z_2))(Z_2(Z_1 + \beta Z_2))^2].$$

By Lemma 7, Jensen's inequality, and Lemma 8,

$$\begin{aligned}
\mathbb{E}[2\phi'(\alpha Z_2(Z_1 + \beta Z_2))(Z_2(Z_1 + \beta Z_2))^2] &= \frac{1}{\alpha^2} \mathbb{E}[2\phi'(\alpha Z_2(Z_1 + \beta Z_2))(\alpha Z_2(Z_1 + \beta Z_2))^2] \\
&\leq \frac{2}{\alpha^2} \mathbb{E}[\sqrt{2(1 - \phi(\alpha Z_2(Z_1 + \beta Z_2)))] \\
&\leq \frac{2}{\alpha^2} \sqrt{\mathbb{E}2(1 - \phi(\alpha Z_2(Z_1 + \beta Z_2)))] \\
&\leq \frac{2}{\alpha^2} (\frac{1}{\beta \min\{\alpha, \beta\} + 1})^{1/4}.
\end{aligned}$$

□

## References

- [1] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Ann. Statist.*, 45(1):77–120, 2017. [1](#), [2](#), [3](#), [4](#), [5](#), [11](#)
- [2] Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a ConvNet with Gaussian inputs. *arXiv Preprint*, February, 2017. [9](#)



- [3] Arun Tejasvi Chaganty and Percy Liang. Spectral experts for estimating mixtures of linear regressions. *In ICML*, (2013):1040–1048, 2013. 2, 5
- [4] Yudong Chen, Xinyang Yi, and Constantine Caramanis. A convex formulation for mixed regression with two components: minimax optimal rates. *In COLT*, pages 560–604, 2014. 2
- [5] Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten steps of EM suffice for mixtures of two Gaussians. *arXiv Preprint*, September, 2016. 2, 4, 8, 14, 15
- [6] Richard D. De Veaux. Mixtures of linear regressions. *Computational Statistics & Data Analysis*, 8(3):227–245, Nov 1989. 2
- [7] Susana Faria and Gilda Soromenho. Fitting mixtures of linear regressions. *J. Stat. Comput. Simul.*, 80(1-2):201–225, 2010. 2
- [8] Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2):181–214, 1994. 2
- [9] Jason M. Klusowski and W. D. Brinda. Statistical guarantees for estimating the centers of a two-component gaussian mixture by EM. *arXiv Preprint*, August, 2016. 2, 9, 14, 15
- [10] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007. 9
- [11] Hanie Sedghi, Majid Janzamin, and Anima Anandkumar. Provable tensor methods for learning mixtures of generalized linear models. *arXiv Preprint*, 2014. 2
- [12] Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151, 1981. 7
- [13] Kert Vele and Barbara Tong. Modeling with mixtures of linear regressions. *Stat. Comput.*, 12(4):315–330, 2002. 2, 5
- [14] Ji Xu, Daniel J. Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two Gaussians. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2676–2684. Curran Associates, Inc., 2016. 2, 4
- [15] Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. *In ICML*, pages 613–621, 2014. 2, 5
- [16] Kai Zhong, Prateek Jain, and Inderjit S. Dhillon. Mixed linear regression with multiple components. *In Advances in Neural Information Processing Systems*, pages 2190–2198, 2016. 2, 5